AI 응용시스템의 이해와 구축

1강. ML 모델의 라이프 사이클

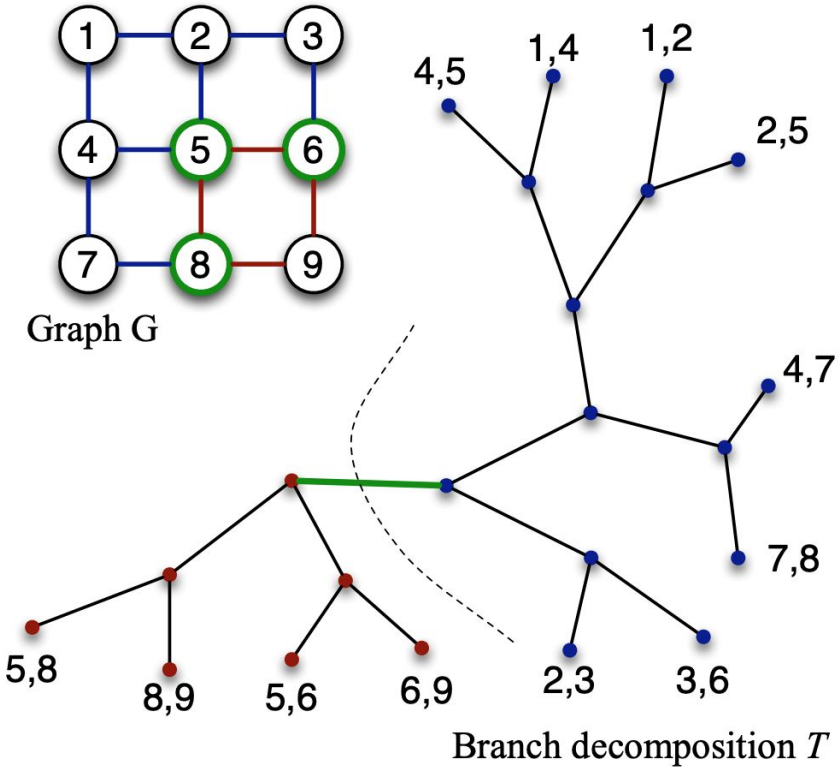
소개

김동현

- 현 구글 Software Engineer / Engineering Manager
  - TensorFlow: 모델최적화 (model optimization). Quantization, Pruning 등 개발.
  - Google Brain / Health: 시계열 의료데이터를 기반으로 clinical / operational prediction, medical notes에서 정형화 데이터 추출.
  - Youtube: Smart TV에서 content recommendation.
  - Search:
    - 구글 나우(Google Now)에서 사용자 의도 예측 (intent prediction model).
    - 개인화 검색 팀에서 question answering (natural language query understanding) 모델.
- 본 전공은 이산수학
  - Combinatorial optimization, computational geometry
  - Bioinformatics, Systems biology

Grad School #1

- Algorithms, Discrete Maths, Computational Geometry, Graph Theory, Combinatorics
  - Lemma, Proof, Theorem, Proof, Corollary, Proof,...
- 코드 1줄도 짜보지 않음 :(



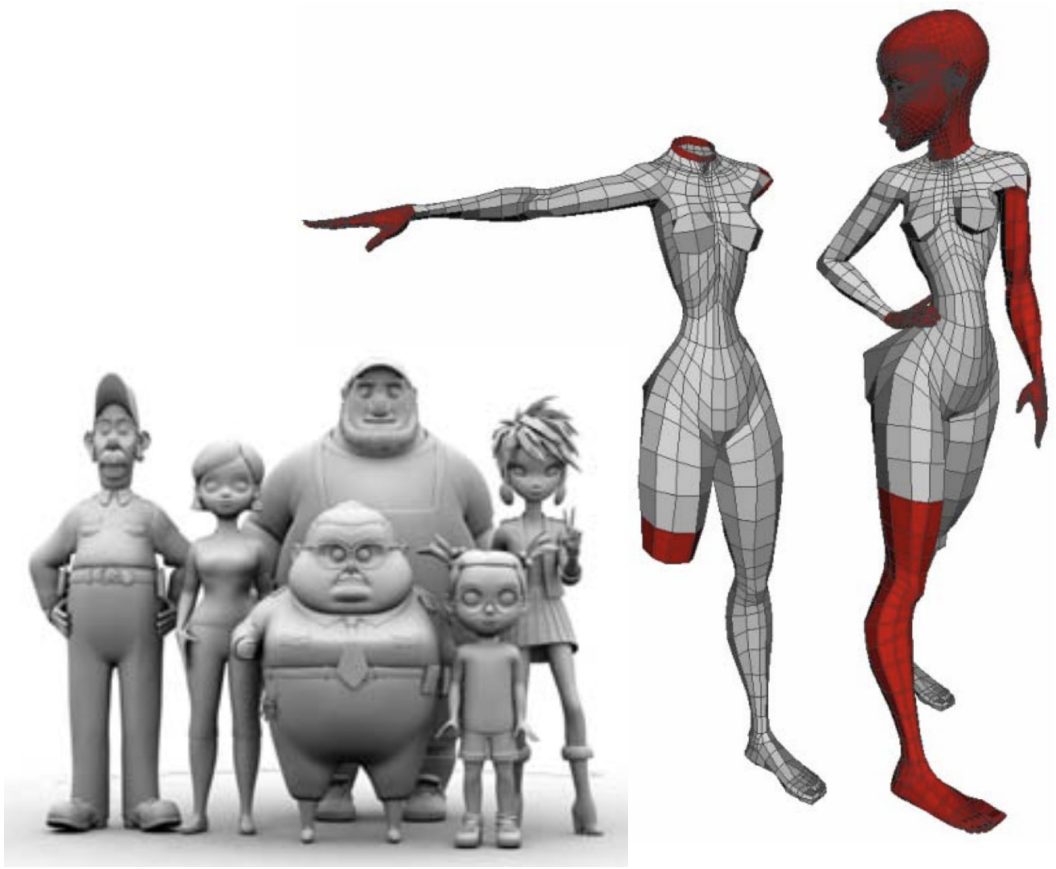
Disney Research

- Graph theory,
  Computational Geometry를 Computer Graphics에 적용.

- Disney Research, Burbank CA
  (당시 Disney Feature Animation)

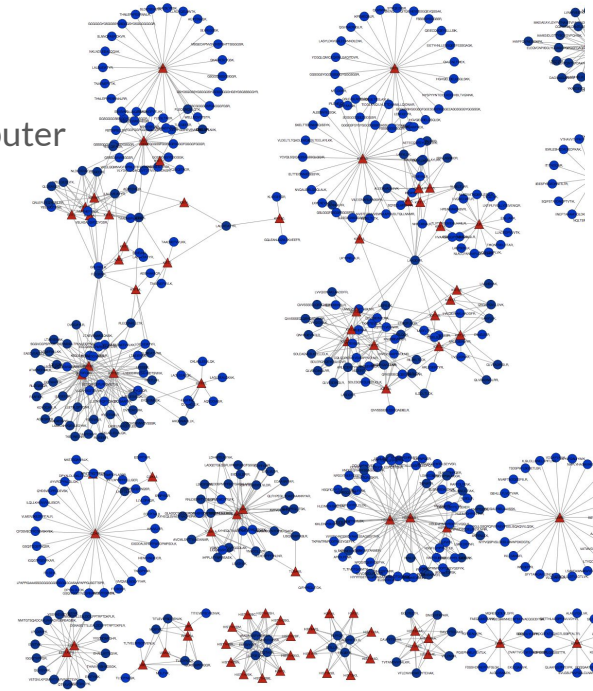
Disney Research

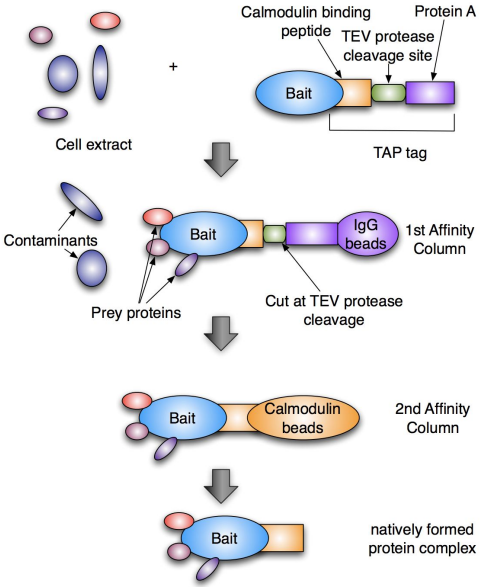
- 코드 20,000줄
- **현실 세계**로 돌아 왔더니 아름답더라..



Grad School #2

- "이제 현실세계로 돌아오고 싶다."
- Bioinformatics = Systems Biology + Computer Science
- 그래프 이론, 이산수학, 머신러닝
- Protein-Protein Interaction Network
    - 단백질간의 상호작용



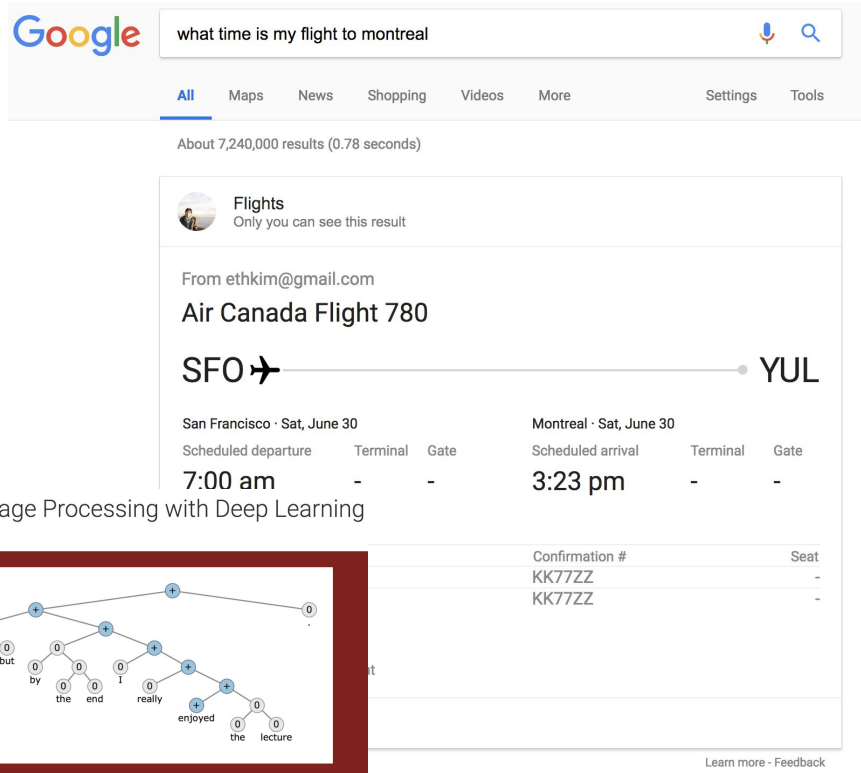


Google, Mountain View

- Software engineer

Google: "Personal Search"

- **자연어 검색어**로 개인 비서에게 하는 말.
  - 항공 스케줄, 호텔/식당 예약, 구매 추적, 리마인더, 캘린더, 이메일, 사진, 드라이브, ...
- "자연어 처리" (natural language processing)
- Google I/O '13 키노트에서 발표.
- 현재 구글 홈 / 어시스턴트.

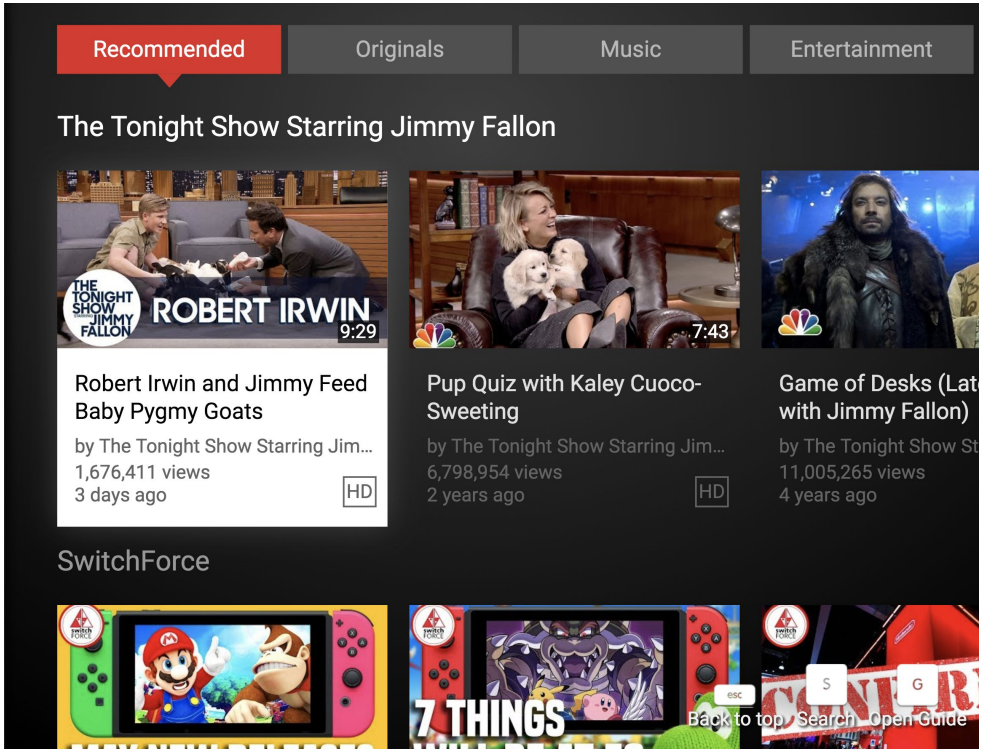


Google Now

- 사용자가 검색을 하지 않아도 필요한 정보를 제공.
- 검색어 X ⇒ 사용자의 context 를 이용.
  - 위치
  - 사용자 프로필
  - 관심사
- 2014 Google I/O 에서 발표

Youtube in the Living Room

- 거실에서 (TV, XBOX, Roku, ...) 유튜브 컨텐츠 추천
- 리모콘으로 검색이 불편하므로 “추천"에 많은 비중.



Google Brain: Medical Applications

(차후 Google Health 사업부로 조정)

환자들의 의료데이터를 사용하여 의료 진단 (clinical / operational outcome)을 예측할 수 있을까?

- Research Blogpost

- 머신러닝 / AI
- 빅데이터
- 자연어 처리
- 컴퓨터 비전

TensorFlow

- 스마트폰에서 쓸 On device 모델을 다 만들었는데 모델 크기가 10GB라면?
- 서버에서 쓸 모델을 학습했더니 inference당 속도가 100ms 라면?

[TensorFlow Model Optimization]

➔ Quantization: floating point model to integer models
➔ Pruning: simplify model graph
➔ and more!

[Google I/O 2020 Session]

여러분 소개

백그라운드:

- 이름, 얼굴 (카메라!)
- 현업 분야
- 인공지능 백그라운드

출석.

Course Logistics

Course Outline

수업	주제
1강	머신러닝 모델의 라이프 사이클
2강	데이터 콜렉션과 피쳐 엔지니어링
3강	데이터 파이프라인
4강	데이터 레이블링 기법
5강	데이터 검증
6강	하이퍼파라메터 튜닝과 모델 아키텍쳐
7강	모델 최적화 기법
8강	**중간고사**

Course Outline

수업	주제
9강	모델 분석 기법
10강	대형 모델 학습
11강	모델서빙
12강	(특강) 프로덕션 머신러닝 시스템의 사례들
13강	모델실험과 디플로이먼트 매니지먼트
14강	로그 시스템과 모델 모니터링
15강	기말평가 프로젝트 발표

코스 스케줄에 따라 정정 가능합니다

Course Evaluation

평가 방법:

- 중간고사: 30%
- 퀴즈: 10% (매 수업마다)
- 참여: 출석 5%, 수업참여도 5%
- 그룹 프로젝트: (4~5명)
  - 과제물: 30%
  - 발표: 20%

Group Project

- 과제물: Engineering Design Doc
  - 팀 멤버들 공통 작성
  - Template + 사례 공지 예정.
  - 구글닥에 코멘트를 활용하여 디자인 토론 + 피드백
  - 학기 중간에 중간 점검 (중간고사 직후)
- 과제물 제출: 14강 수업 직전.
  - 구글닥 코멘트 + 버전 히스토리
  - Bonus Point: 실제 시스템 구축
- 발표: 학기 마지막날 발표
- 졸업 프로젝트로 추가 연장?

수업 비대면 vs 대면

1강 비대면

2강 대면 vs 비대면?

나머지 수업들 (타 강의들과 조율)

Brief Historical Background

History of AI (and failures..)

(1950s)

- 20년내로 인간이 할 수 있는 모든 작업을 기계가 할수 있게 될것이다. (**Herbert Simon**)

- 10년 내로, 인공지능의 대부분의 문제들은 풀릴것이다. (**Marvin Minsky**)

- 로봇에게 인간은, 인간이 강아지같은 위치가 될 세상이 올 것이다. (**Claude Shannon**)

AI Winter

(Machine Translation)

The spirit is willing but the flesh is weak.

(Russian)

The vodka is good but the meat is rotten.

More Examples of Failures

Apple FaceID 시스템에서 3D printing된 얼굴로 해킹	**(Model Drift) Feature 데이터 부족**
Amazon 채용 시스템에서 resume screening 오류: 남성 지원자에게 bias	**ML Fairness (Label 분포 오류)**
Google Photos에서 흑인 사진을 고릴라로 판정	**ML Fairness (Label 분포 오류)**
헬스케어 시스템에서 특정 약을 투여하도록 추천. 왜?	**Explainability**

ML Modeling

AI/ ML expert가 되려면 어떤 전공의 공부를 해야 하나요?

ML 모델링은 오랜 기간동안 여러 분야에서의 아이디어들이 모여져서 생성된 분야. ("a melting pot of many fields").

- Bayes rule (Bayes, 1763), probability
- Least squares regression (Gauss, 1795), astronomy
- First-order logic (Frege, 1893), logic
- Maximum likelihood (Fisher, 1922), statistics
- Neural networks (McCulloh/Pitts, 1943), neuroscience
- Stochastic gradient descent (Robbins/Mono, 1951), optimization
- Uniform cost search (Dijkstra, 1956), algorithms

ML Engineering

ML 엔지니어링 / ML 인프라도 여러 분야들에서 시작된 기술들의 집합.

- 모델 드리프트 (Train / Serving skew): statistics
- 모델 양자화 (quantization): information theory / signal processing
- 모델 프루닝 (Pruning): graph theory
- 데이터 파이프라인: distributed computing
- 모델 아키텍쳐 써치: genetic algorithm, bayesian optimization, ..

ML Engineering (MLOps)의 커리큘럼은 여전히 정의되어가고 있는 중.

(Industry faster than Academia)

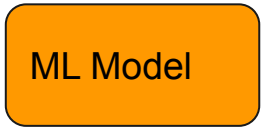
Quiz Time + Break Time

10 mins

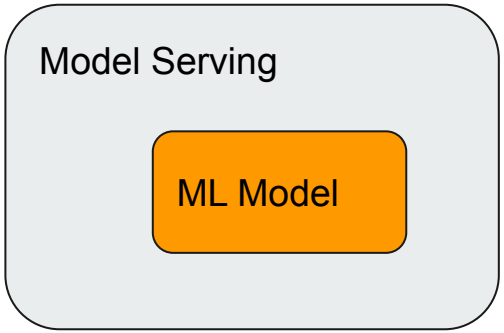
Quiz 1.

Lifecycle of ML System

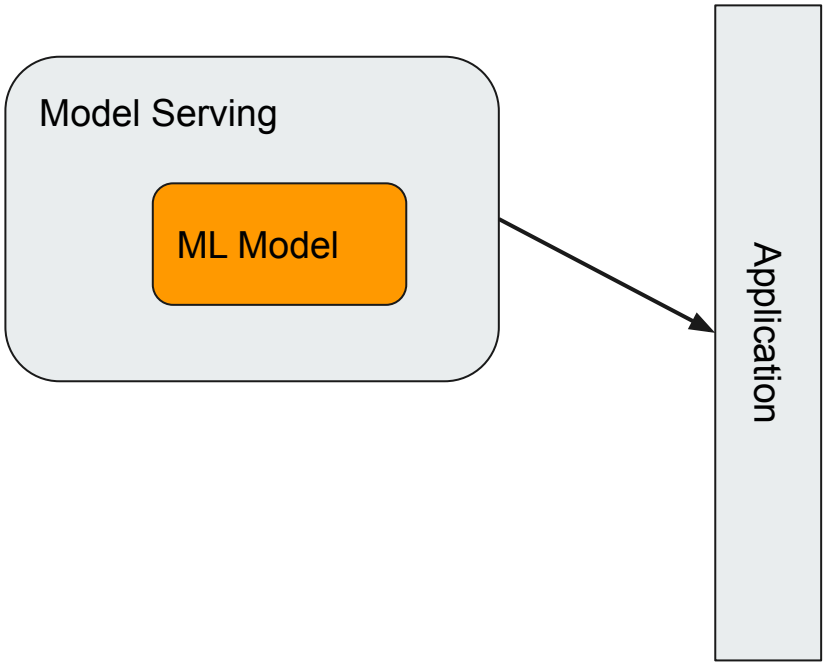
ML models in Production



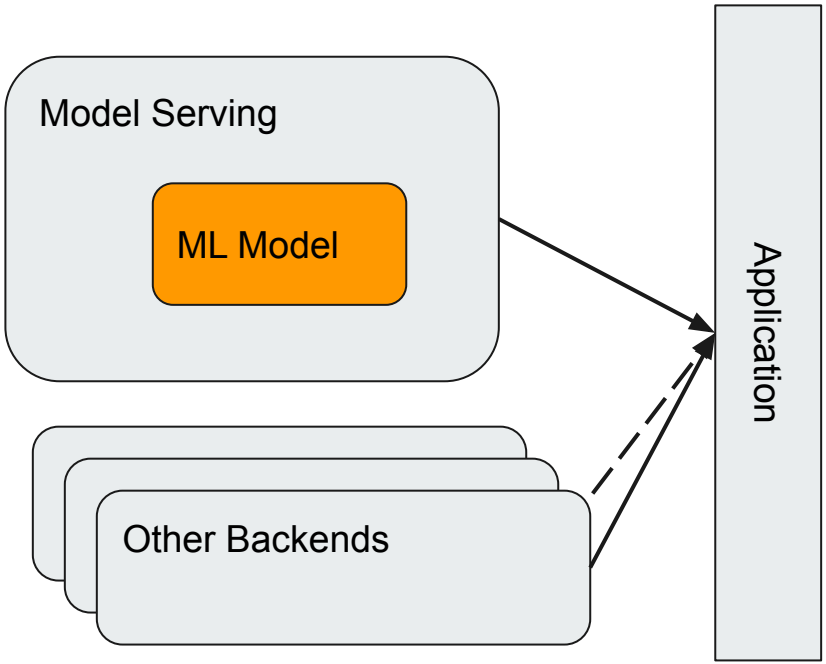
ML models in Production



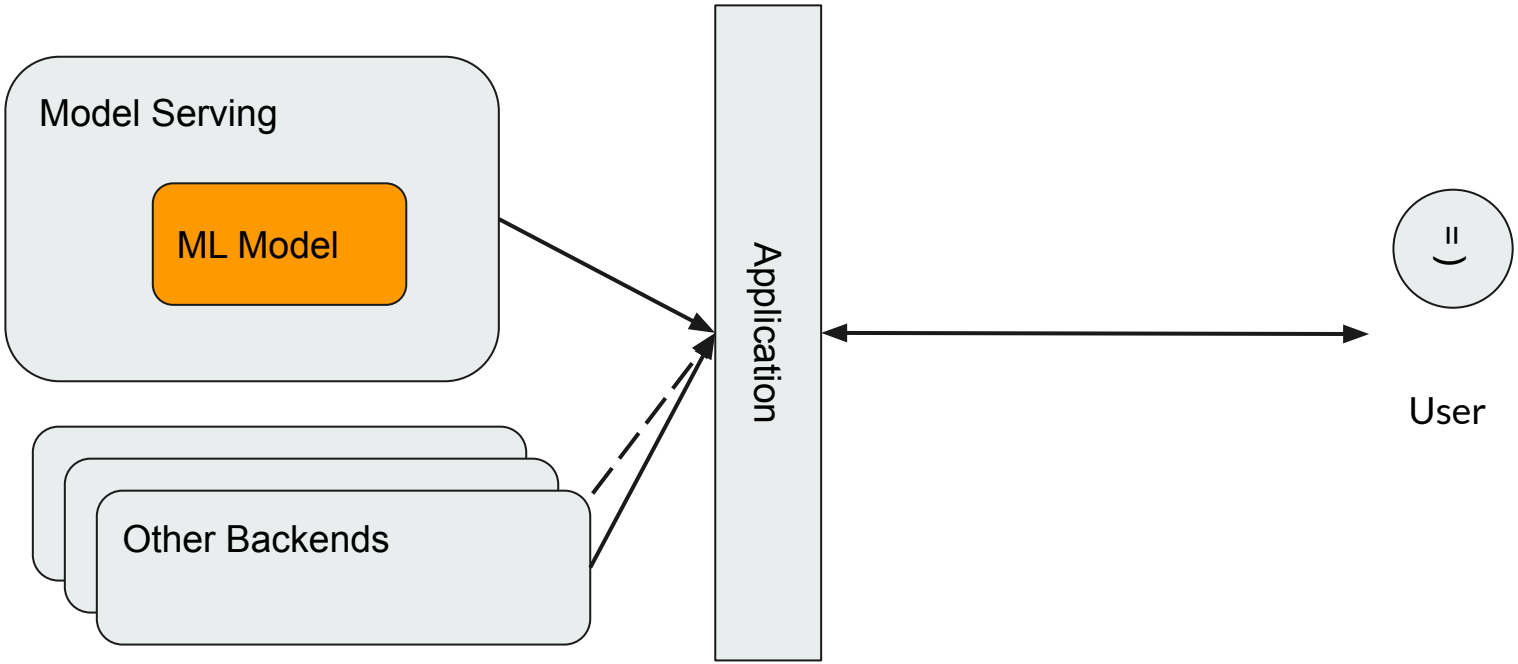
ML models in Production



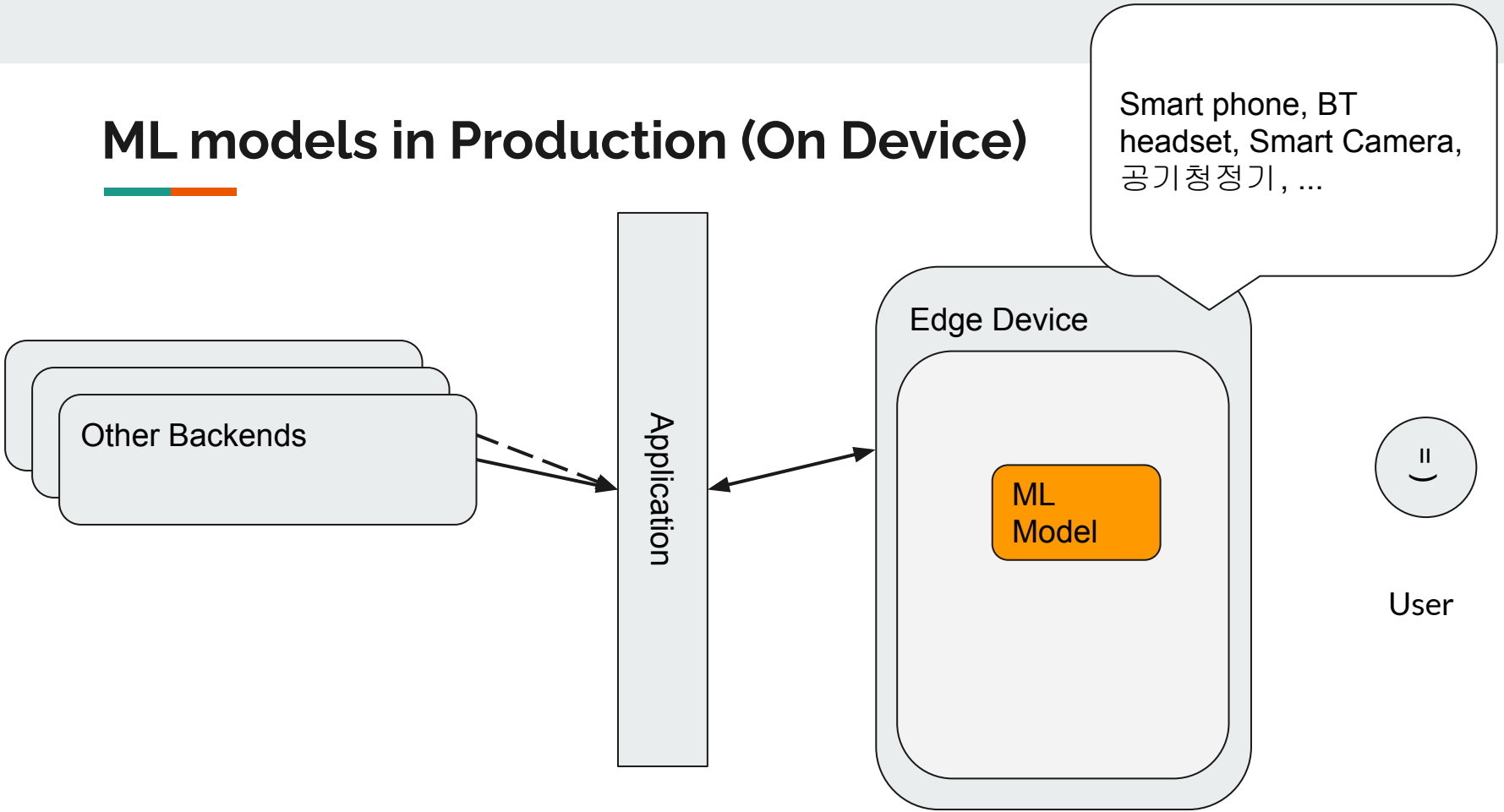
ML models in Production



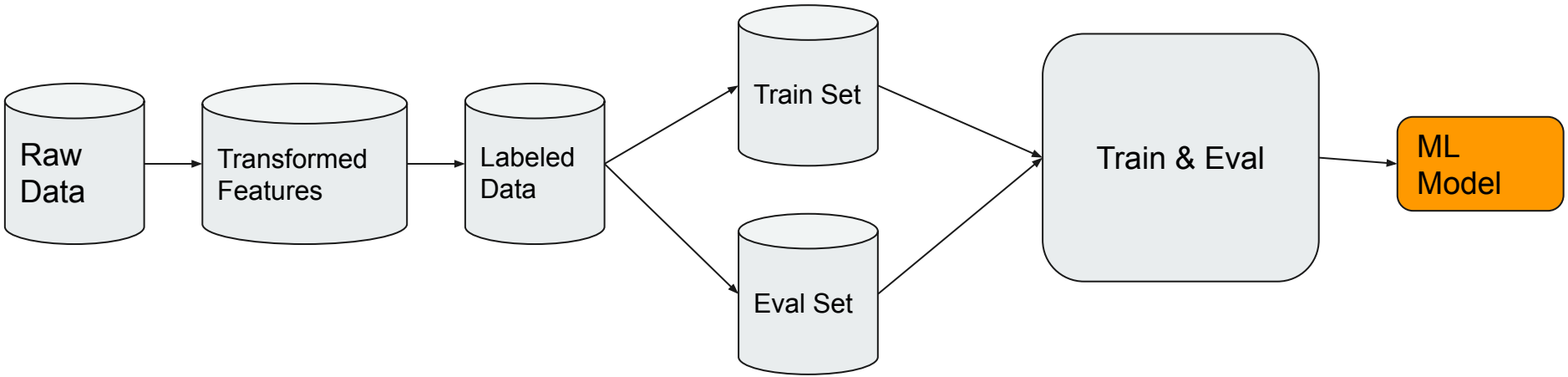
ML models in Production



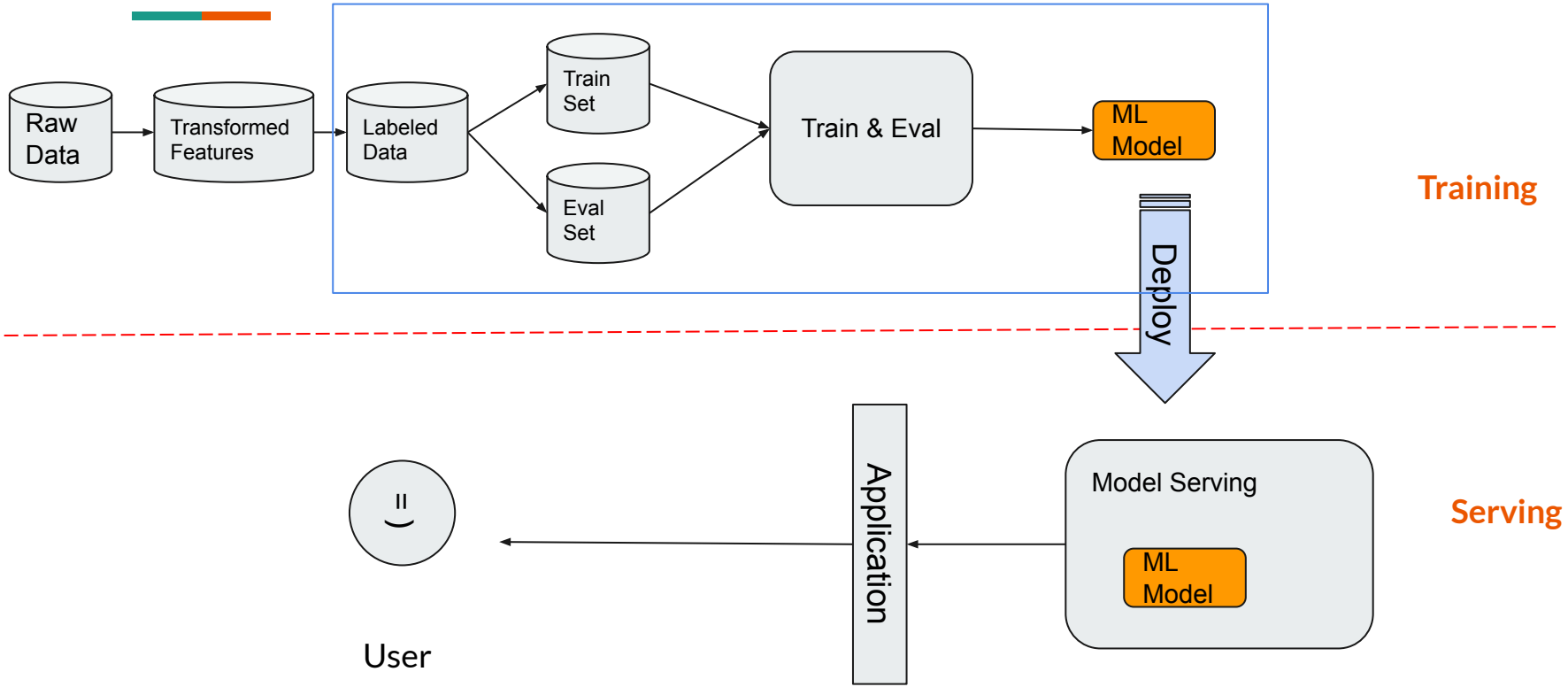
ML models in Production (On Device)



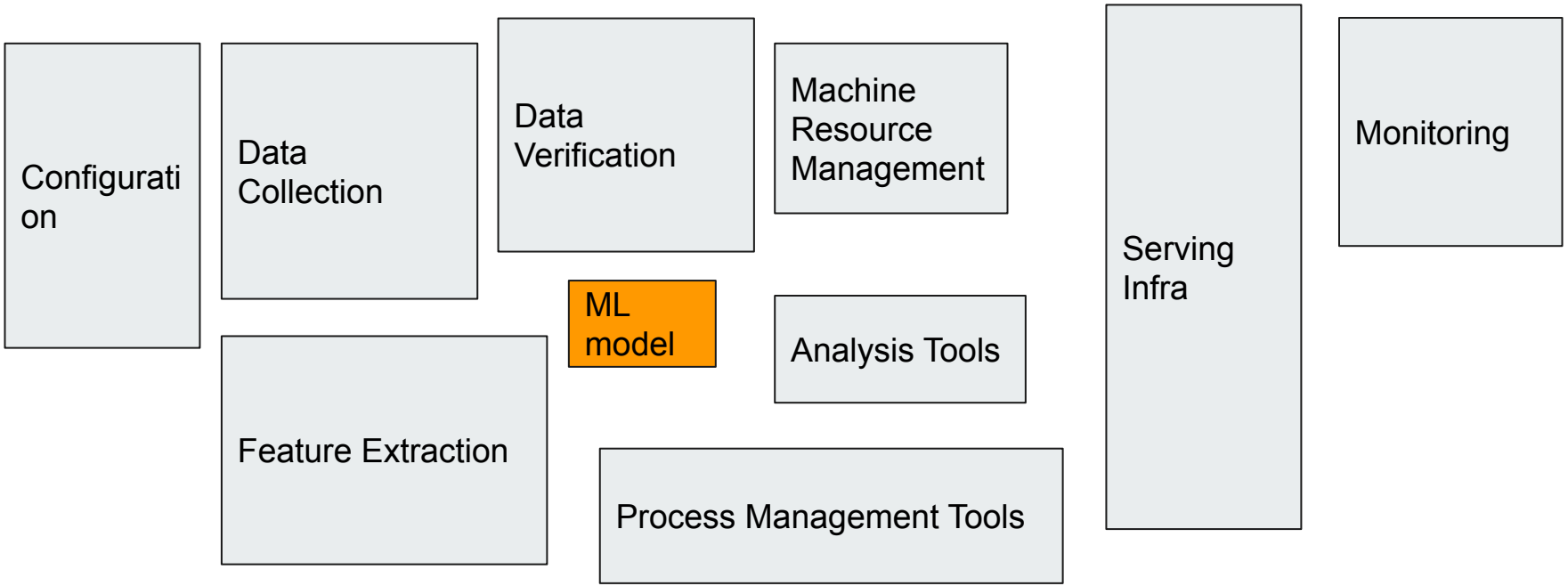
ML models in Production (Training)



ML models in Production (Train & Serving)

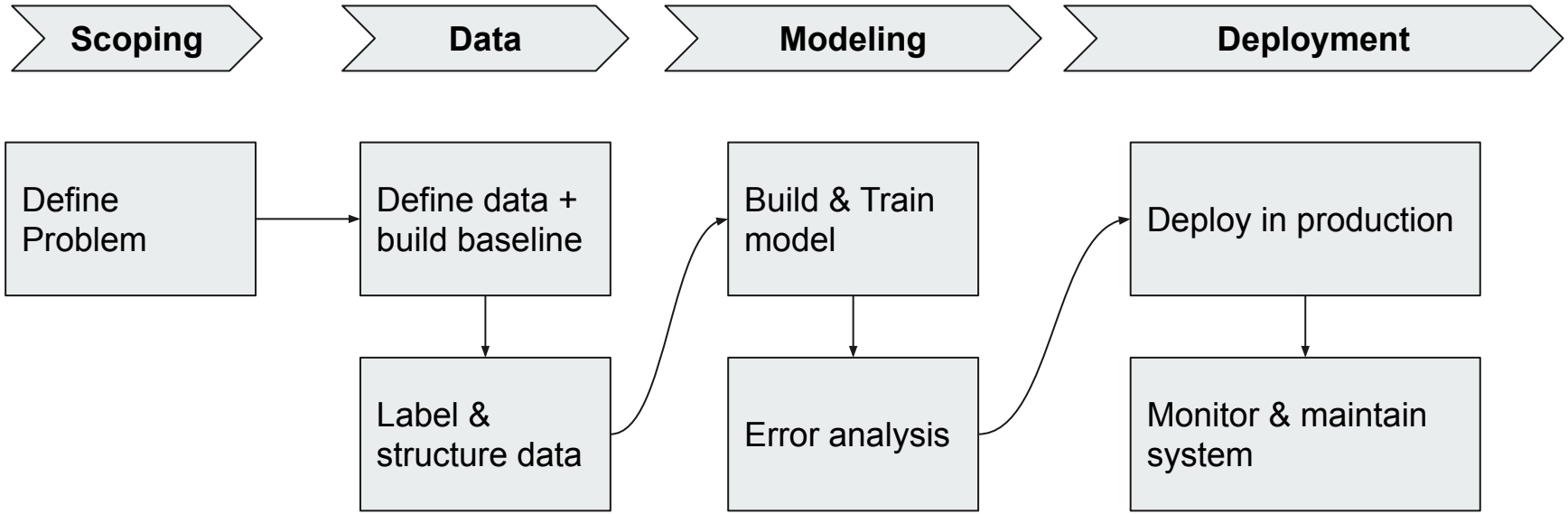


Components of ML Infrastructure



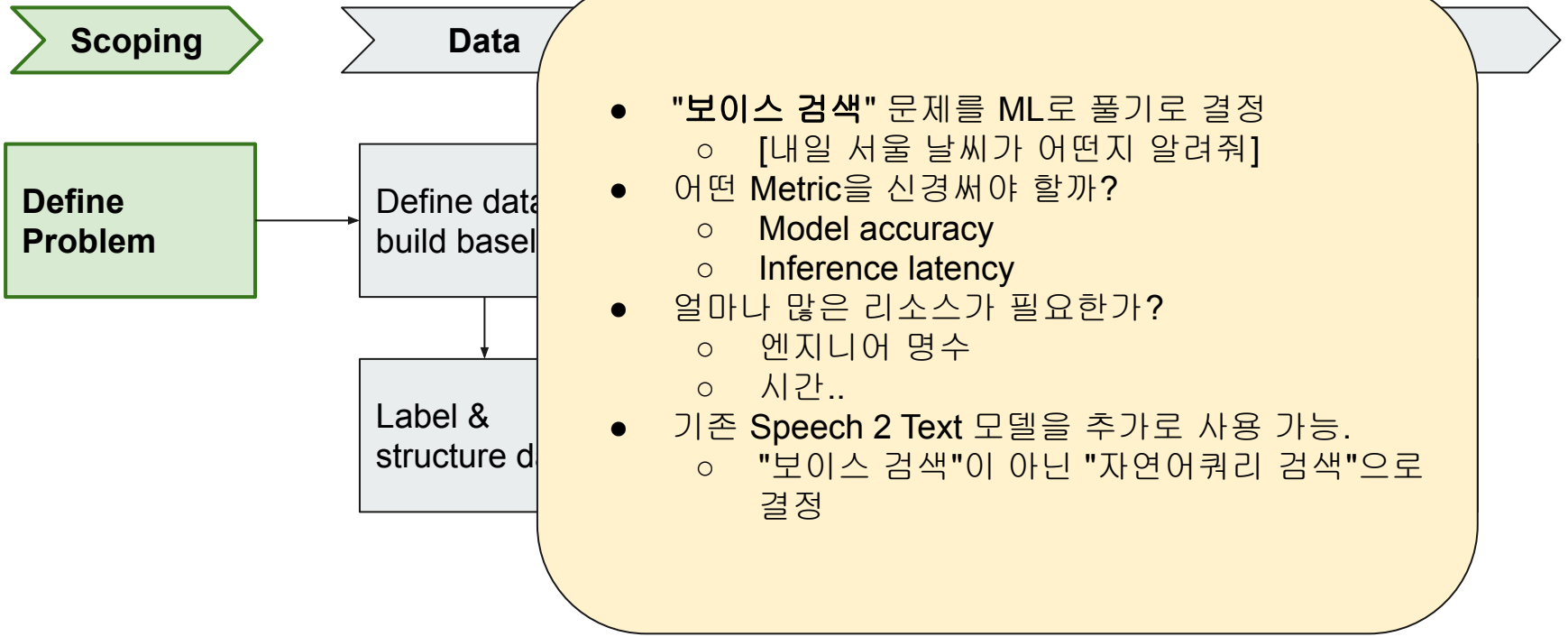
Tiny fraction of real-world ML is actual ML model code.
"Hidden Technical Debt in Machine Learning Systems", D. Sculley et al. (NIPS 2015)

Overall ML Project Lifecycle

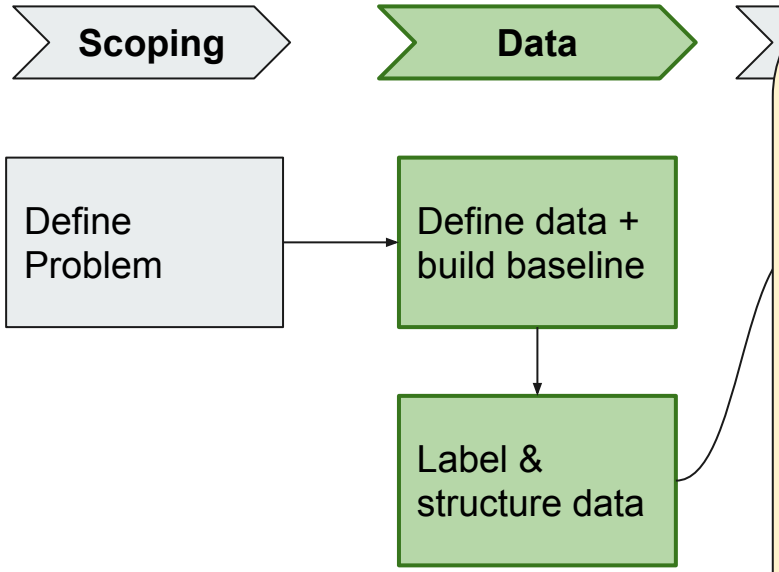


DeepLearning.AI (Machine Learning in Production)

Overall ML Project Lifecycle



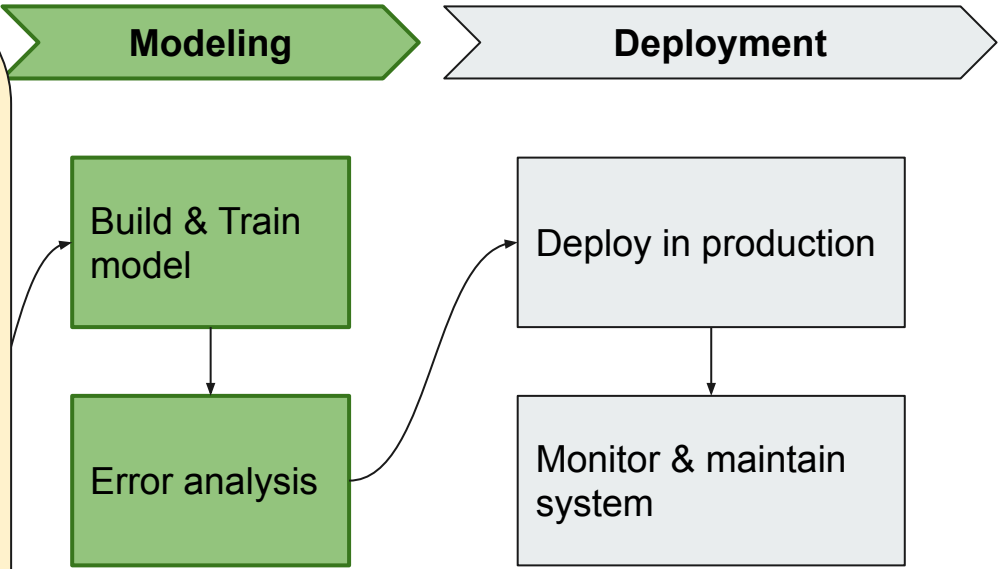
Overall ML Project Lifecycle



- 데이터 정리
  - raw **query** data:
    - [내일 날씨를 알려줘]
    - [오늘 뉴스를 들려줘]
    - [재즈 음악 틀어줘]
- 레이블(Label)과 피쳐셋(Feature Set)
  - Label?
  - Features?
- Baseline (기초선) 모델
  - if "날씨" in query:
    - trigger_weather()
  - elif "뉴스" in query: ...

Overall ML Project Lifecycle

- 모델링:
  - 모델 아키텍쳐를 선택 (DNN, LSTM, Transformer
  - 하이퍼파라메터를 선택 (# of layers, feature selection, ..)
  - 데이터를 선택 (entire data? subsampling?)
- 에러 분석:
  - 특정 데이터셋에 좋은 성능을 내는지?
  - ML Fairness
  - 추가적인 데이터 콜렉션이 필요한지?

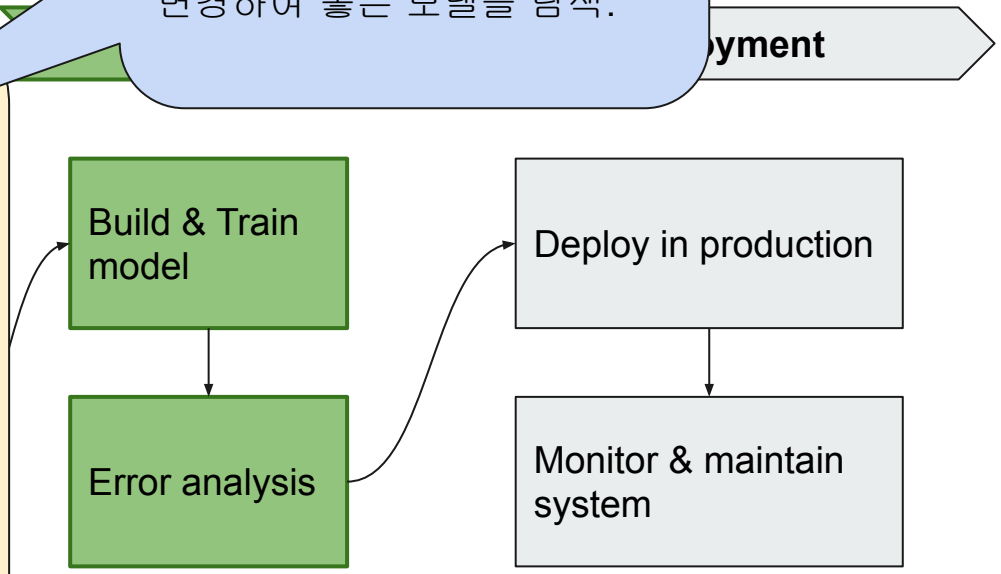


Overall ML Project Lifecyc

ML Research (아카데미아):

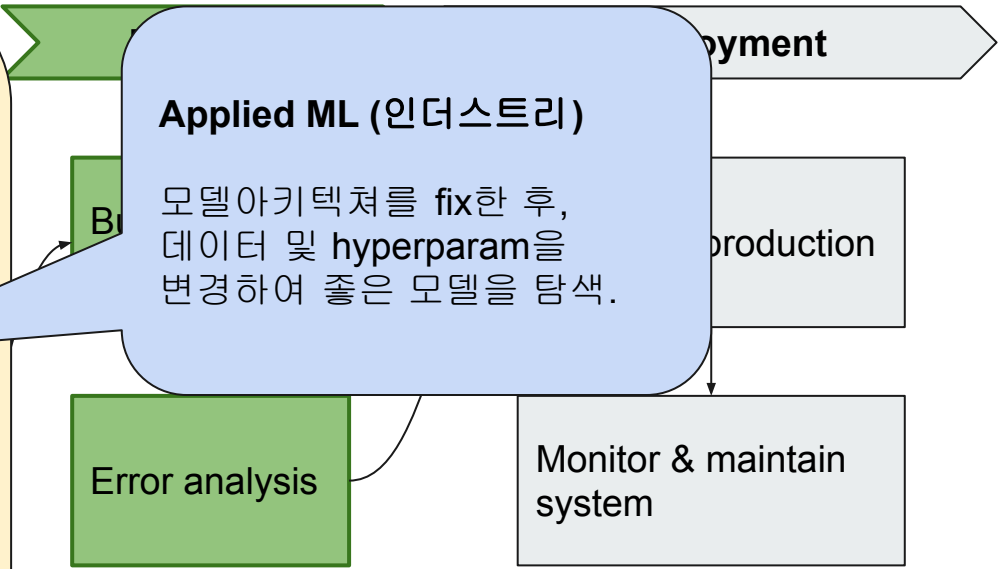
Data를 fix한 후,
모델아키텍쳐+hyperparam을
변경하여 좋은 모델을 탐색.

- 모델링:
  - **모델 아키텍쳐를 선택 (DNN, LSTM, Transformer**
  - **하이퍼파라메터를 선택 (# of layers, feature selection, ..)**
  - 데이터를 선택 (entire data? subsampling?)
- 에러 분석:
  - 특정 데이터셋에 좋은 성능을 내는지?
  - ML Fairness
  - 추가적인 데이터 콜렉션이 필요한지?

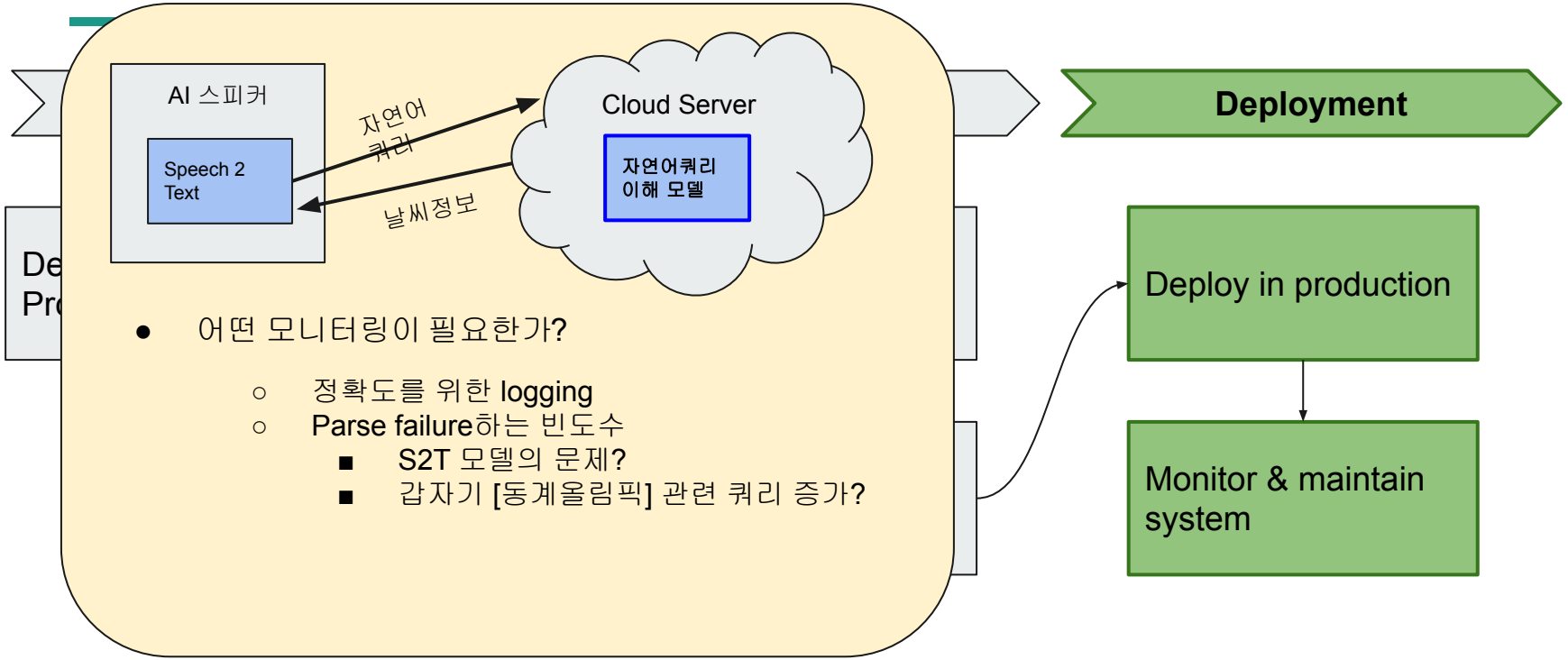


Overall ML Project Lifecycle

- 모델링:
  - 모델 아키텍쳐를 선택 (DNN, LSTM, Transformer
  - **하이퍼파라메터를 선택 (# of layers, feature selection, ..)**
  - **데이터를 선택 (entire data? subsampling?)**
- 에러 분석:
  - 특정 데이터셋에 좋은 성능을 내는지?
  - ML Fairness
  - 추가적인 데이터 콜렉션이 필요한지?



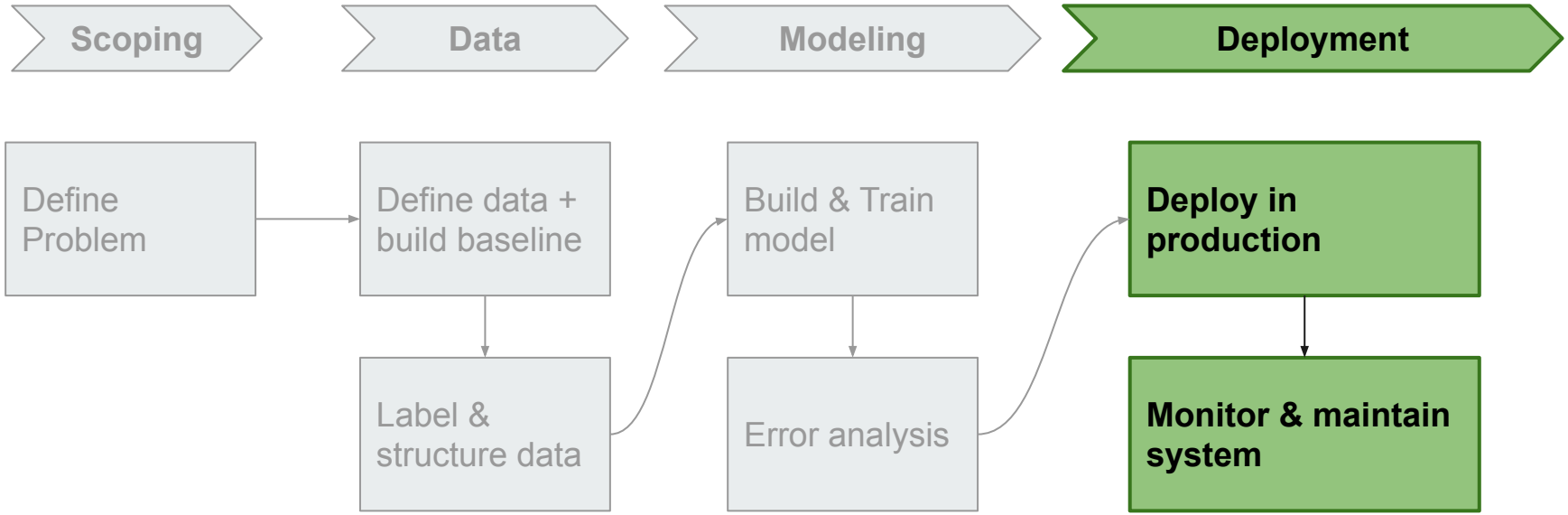
Overall ML Project Lifecycle



- 어떤 모니터링이 필요한가?
  - 정확도를 위한 logging
  - Parse failure하는 빈도수
    - S2T 모델의 문제?
    - 갑자기 [동계올림픽] 관련 쿼리 증가?

Deployment Problems

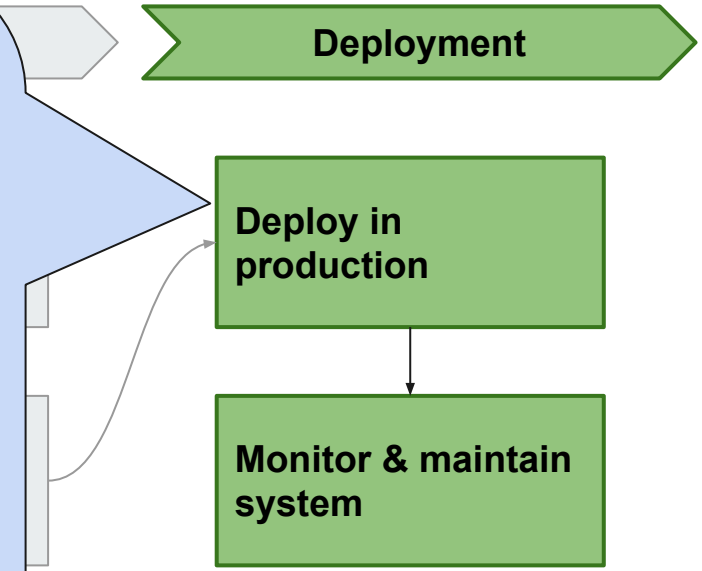
Deployment


Scoping
Data
Modeling
Deployment
Define Problem
Define data + build baseline
Label & structure data
Build & Train model
Error analysis
Deploy in production
Monitor & maintain system

Deployment

"Software Design Choice"

- 실시간(Realtime) vs 배치(Batch)
- 서버(Cloud) vs 온디바이스 (Edge, Browser)
- Inference 플랫폼 (CPU, GPU/TPU, memory)
- Latency, QPS
- 로깅(Logging)
- Security / Privacy

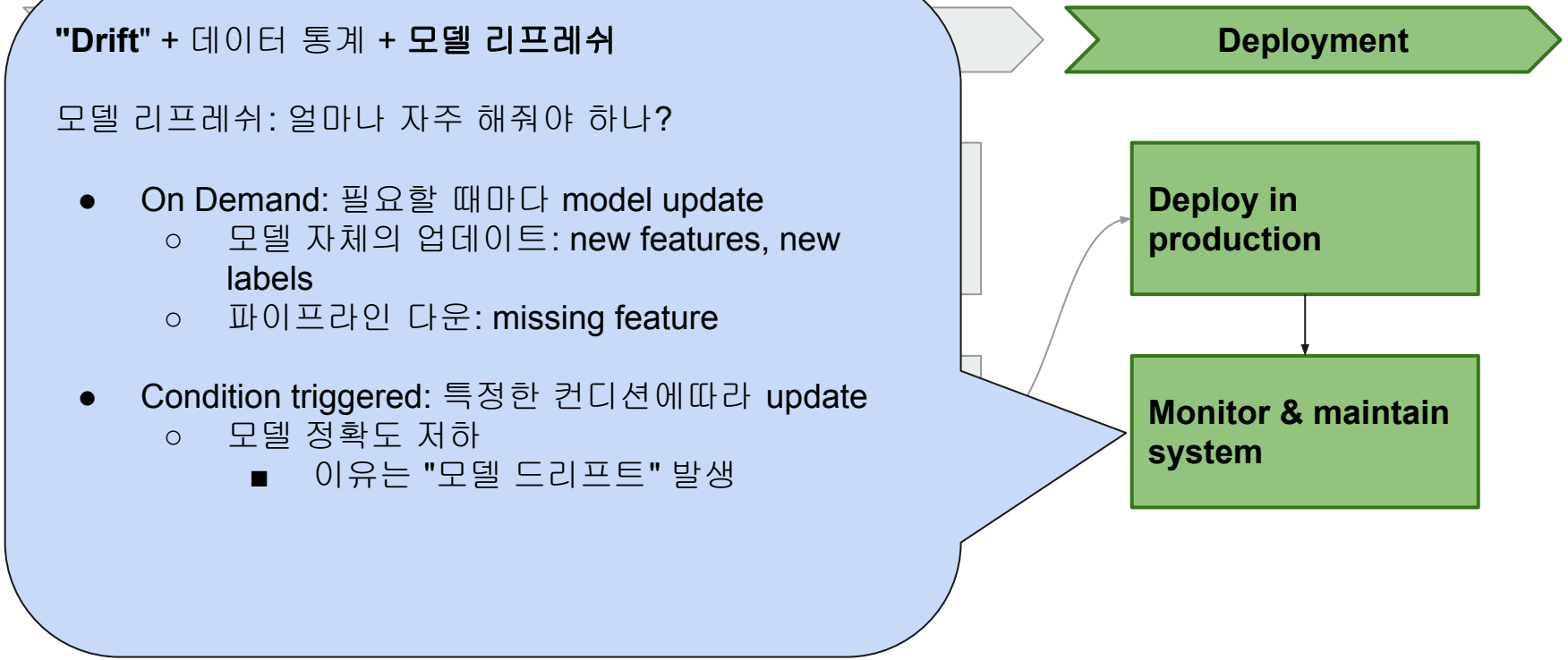


Deployment

"Drift" + 데이터 통계 + **모델 리프레쉬**

모델 리프레쉬: 얼마나 자주 해줘야 하나?

- On Demand: 필요할 때마다 model update
  - 모델 자체의 업데이트: new features, new labels
  - 파이프라인 다운: missing feature
- Condition triggered: 특정한 컨디션에따라 update
  - 모델 정확도 저하
    - 이유는 "모델 드리프트" 발생



Model Drift (모델 드리프트)

모델이 학습(train) 되었을 때 사용된 데이터셋과, 실제 추론(inference) 시 쓰이는 데이터셋이 달라질 수 있다. 이때 **training-serving skew**가 발생하게 되어 모델의 정확도가 저하된다.

- **Concept Drift**: 모델이 예측하고 싶은 변수 (종속변수, dependent variable / **label**) 의 변경.
    - 신용카드 회사에서 고객의 상품구매 패턴에 따라 Fraud detection 을 하는 모델을 개발.
    - Label을 정의할 때:
        - 어떤 고객 데이터로 fraud 인지 정의할까?
            - 지난달 대비 오프라인 구매가 다른 국가에서 발생 -> fraud
            - **온라인 쇼핑 구매액이 지난달 대비 200% 이상 증가 -> fraud**
        - 코로나 이후 온라인구매 대폭 증가 -> Label의 재정의 필요.
- **Data Drift**: 모델이 사용하는 변수(독립변수, independent variable / **features**) 의 변경.
    - Seasonality (계절주기)에 따른 데이터 변화
    - 특정 feature가 더이상 발생하지 않는 경우
    - Class imbalance: 특정 demographic이나 label 분포가 변경

Drift (모델 드리프트)

Modeling Problem: (예제)

- 각 대형병원의 응급실에는 정해진 침상의 갯수(n)가 존재
- 앰뷸런스가 응급환자를 이동할 때, 가까운 거리의 응급실들 중 환자를 수용할 수 있는 곳으로 이동.
- 따라서, 각 대형병원에서는 10~30분 이내 수용가능한 침상을 예측할 필요.
- 평균적으로 응급실환자는 36시간 이내 병동으로 이동.

Label:

- True if {남은 침상 갯수 + 입원후 36시간된 환자 명수} > 0

Feature:

- 요일 (DayOfWeek), 시간(Hour), 월, 일

Training Frequency:

- 3~4 개월에 한번? 매달?

Drift (모델 드리프트)

Data Drift

- 코로나 이후 입원환자는 **요일과 무관**
- 확진자수 패턴

Concept Drift

- 일반적인 응급실 환자는 36시간 이내 해당병동으로 이동
- 코로나 이후, 타 병동에서 응급실환자 수용으로 사용가능한 침상수 증가.

When does drift occur?

- 천천히 순차적인 변화 (Gradual data change):
  - User data는 "일반적으로" 천천히 변화.
  - Seasonality: 계절에 따라 사용자들의 쇼핑 패턴 변화

- 갑작스런 변화 (Sudden **shock**):
  - B2B 데이터는 종종 빠르게 변화가능.
  - 의존하는 피쳐 데이터 소스의 변화 혹은 **삭제**:
    - 주소 지번 변경 (옛주소로 학습, 새주소로 추론)
    - 날씨 데이터
  - 자연재해: 코로나 바이러스, 태풍 / 지진

How to detect drift?

1. **천천히 변화** (Gradual data change):
   수동적으로 모델 업데이트 정책(model update policy)을 이용.
   a. 예상되는 변화가 있다면 → Seasonality에 따라 모델 업데이트
   b. 피쳐 데이터에 Seasonality를 빌트인 (feature engineering)

2. **갑작스러운 변화**:
   자동적인 detect & update 시스템이 필요.

How to detect drift?

Sequential Drift Detection

- DDM / EDDM: (Early) Drift Detection Method:
  a. 모델의 에러 갯수를 binomial variable로 표현.
  b. Expected number of errors in time window X < some standard deviation

Data Monitoring

- 통계적 분포 (Statistical Distribution Properties)
  - 주어진 두개의 데이터셋 간의 거리 측정
  - KL-Divergence
  - Total Variation Distance
  - Hellinger Distance

얼마나 자주 업데이트 할수 있나?
예) 모델 학습이 2주걸리면, 모델 업데이트 빈도가 2주?

Deployment Strategy

Deployment 시나리오:

1. 완전히 새로운 제품 / 기능
2. 기존에 있던 baseline ML system을 대체
3. 인간이 하던 매뉴얼 작업을 도와주거나 자동화

주의해야 할 점:

- Gradual "ramp up": 조금씩 실험에 돌리면서 모니터링
- "Rollback": 실험 결과가 좋지 않다면 "ramp down"이 필요

Deployment Strategy

예: 공장 출하 중에 불량제품 판정하기

1. 인간이 일일히 모든 제품을 육안으로 확인

2. Computer Vision 모델이 불량을 판정 → 인간이 재확인하며 처

모델의 정확도 데이터 수집.

3. 작은 양(5%)은 ML 모델이, 나머지는 인간이 판정.

4. ML 모델이 판정하는 양을 조금씩 증가.

기존의 모델(baseline)을 새 모델로 대체할땐?

baseline을 인간에 비교

이러한 deployment 를 **"Canary deployment"** 이라고 칭함.

- 탄광에 gas leak을 감지하기위해 사용